When are principal component scores a good tool for visualizing high-dimensional data?

Kristoffer H. Hellton

(joint work with Magne Thoresen)

Principal component analysis (PCA) is a popular method for visualizing and reducing the dimension of high-dimensional data. In particular in genetic applications, the initial data exploration is often done by visually investigating the first principal component (PC) scores.

Let $X = [x_1, \ldots, x_n]$ be a $p \times n$ data matrix, where $x_i = [x_{i1}, \ldots, x_{ip}]^T$ are independent and identically distributed with $\mathbb{E}x_i = 0$ and $\text{var}x_i = \Sigma$. Then the eigenvectors, $v_1, \ldots, v_p$, of the population covariance matrix $\Sigma$ define the population PCs as

$$s_j^T = v_j^T X = [v_j^T x_1, \ldots, v_j^T x_n],$$

such that the PCs are linear and orthogonal combinations of variables expressing maximal variability. As the variance of each PC is given by the corresponding eigenvalue, $\lambda_1 \geq \cdots \geq \lambda_p$, the standardized PC is defined as $z_j^T = v_j^T X / \sqrt{\lambda_j}$. The sample PCs are based on the eigenvalues and -vectors of the sample covariance matrix $\hat{\Sigma}$, denoted by $d_1, \ldots, d_p$ and $\hat{v}_1, \ldots, \hat{v}_p$, such that the sample scores and standardized sample scores are given

$$\hat{s}_j^T = \hat{v}_j^T X, \quad \hat{z}_j^T = \frac{v_j^T X}{\sqrt{d_j}},$$

In the classical asymptotic setting ($p$ is fixed and $n \to \infty$) the sample eigenvalues and eigenvectors will be consistent estimators for the population eigenvalues and eigenvectors. In the high-dimensional case, however, Paul (2007) established within the random matrix framework where $p/n \to \gamma > 0$ as $p, n \to \infty$, that the sample eigenvalues and eigenvectors are in fact inconsistent. The same holds within the High Dimension Low Sample Size (HDLSS) framework (Jung and Marro, 2009; Jung et al., 2012), where $n$ is fixed and the $m$ first eigenvalues scale with the dimension, $\lambda_i = \sigma_i^2 n^{\alpha}$. Then for the special case of $\alpha = 1$, it is possible to show that relative positions of the PC scores, and thereby the visual information, can be consistent even if the eigenvectors are not (Hellton and Thoresen, 2014). This offers an explanation for the paradoxical situation where classical PCA works successfully in practice, despite being theoretically inconsistent. The key assumption of $\alpha = 1$ can be interpreted in terms of the generating mechanism behind the data.

Let $x_i \sim N(0, \Sigma)$ be i.i.d. and let the $m$ first eigenvalues scale with $p$

$$\lambda_1 = \sigma_1^2 p, \quad \ldots, \quad \lambda_m = \sigma_m^2 p.$$
Figure 1. Plots of sample PC scores (black dots) compared to population PC scores (circles) showing scaling and rotation.

while the remaining eigenvalues are constant $\lambda_m = \cdots = \lambda_p = \tau^2$. Then the vector of the $m$ first sample scores of the $i$th observation converges in distribution, as $p \to \infty$, to

$$
\begin{bmatrix}
\hat{z}_{i1} \\
\vdots \\
\hat{z}_{im}
\end{bmatrix}
\stackrel{d}{\to}
\begin{bmatrix}
\sqrt{n/d_1} & 0 \\
0 & \sqrt{n/d_m}
\end{bmatrix}
\begin{bmatrix}
\sigma_1 z_{i1} \\
\vdots \\
\sigma_m z_{im}
\end{bmatrix},
$$

for $i = 1, \ldots, n$, where $d_j$ and $u_j$ are the $j$th eigenvalue and -vector of an $m \times m$ Wishart distributed matrix, $W \sim \text{Wishart}(n, \text{diag}(\sigma_1^2, \ldots, \sigma_m^2))$. The consequence is that the vector of the first $m$ sample scores will be a rotated and scaled version of the population scores in $m$ dimensions, and the relative positions of the scores will remain the same.

For the purpose of visualizing data one would plot pairs of the $m$ first sample scores in two dimensions. Simulations based on normally distributed data show that for moderate sample sizes, a two-dimensional representation of the sample scores will also be a scaled and approximately rotated version of the population scores. Thus the visual information of the population scores is preserved by the sample scores. The behavior is demonstrated in Figure 1 showing two realizations of the sample scores compared to the population scores.

The assumption of linearly increasing eigenvalues can be motivated by viewing the principal components as latent factors. If a factor is pervasive, affecting most or a significant proportion of the observed variables, the corresponding population eigenvalue will scale linearly with the dimension asymptotically. Some applications where pervasive effects are reasonable include genetics markers (SNPs), where different populations and ethnicity act as latent factors, and stock returns with supply and demand shocks.
REFERENCES


